Deep Learning - Week 2

1. Which of the following statements is(are) true about the following function?
$\sigma(z) = \frac{1}{1+e^{-(z)}}$

   (a) The function is monotonic

   (b) The function is continuously differentiable

   (c) The function is bounded between 0 and 1

   (d) The function attains its maximum when $z \to \infty$

   Correct Answer: (a),(b),(c),(d)
   Solution: Plot the function on graphing tools and verify all the correct options

2. How many weights does a neural network have if it consists of an input layer with 2 neurons, two hidden layers each with 5 neurons, and an output layer with 2 neurons? Assume there are no bias terms in the network.

   Correct Answer: 45
   Solution: Number of weights $= (2 * 5) + (5 * 5) + (5 * 2) = 45$.

3. A function $f(x)$ is approximated using 100 tower functions. What is the minimum number of neurons required to construct the network that approximates the function?

   (a) 99

   (b) 100

   (c) 101

   (d) 200

   (e) 201

   (f) 251

   Correct Answer: (e)
   Solution: To approximate one rectangle, we need 2 neurons. Therefore, to create 100 towers, we will require 200 neurons. An additional neuron is required for aggregation.

4. Suppose we have a Multi-layer Perceptron with an input layer, one hidden layer and an output layer. The hidden layer contains 32 perceptrons. The output layer contains one perceptron. Choose the statement(s) that are true about the network.

   (a) Each perceptron in the hidden layer can take in only 32 Boolean inputs

   (b) Each perceptron in the hidden layer can take in only 5 Boolean inputs

   (c) The network is capable of implementing 2^5 Boolean functions

   (d) The network is capable of implementing 2^{32} Boolean functions

Correct Answer: (d)
Solution: In the lecture, we have seen that, if the hidden layer contains 2^n neurons, where n is a number of inputs, then the network should be able to implement all Boolean functions that take in n inputs. There are 2^{2^n} Boolean functions.

5. Consider a function $f(x) = x^3 - 5x^2 + 5$. What is the updated value of x after 2nd iteration of the gradient descent update, if the learning rate is 0.1 and the initial value of x is 5?

Correct Answer: range(3.1,3.2)
Solution: We are tasked to find the updated value of x after the second iteration of gradient descent for the function:

$$f(x) = x^3 - 5x^2 + 5$$

Step 1: Compute the Gradient The gradient of $f(x)$ is given by:

$$f'(x) = 3x^2 - 10x$$

Step 2: Gradient Descent Update Rule The update rule for gradient descent is:

$$x_{\text{new}} = x_{\text{old}} - \eta \cdot f'(x_{\text{old}})$$

where η is the learning rate.

Step 3: Initial Parameters

- Initial $x = 5$
- Learning rate $\eta = 0.1$

Step 4: Iteration 1 At $x = 5$:

$$f'(5) = 3(5)^2 - 10(5) = 75 - 50 = 25$$

Update x:

$$x_{\text{new}} = 5 - 0.1 \cdot 25 = 5 - 2.5 = 2.5$$

Step 5: Iteration 2 At $x = 2.5$:

$$f'(2.5) = 3(2.5)^2 - 10(2.5) = 3(6.25) - 25 = 18.75 - 25 = -6.25$$

Update x:

$$x_{\text{new}} = 2.5 - 0.1 \cdot (-6.25) = 2.5 + 0.625 = 3.125$$

Final Answer The updated value of x after the second iteration is:

$$x = 3.125$$

6. Consider the sigmoid function $\frac{1}{1+e^{-(wx+b)}}$, where w is a positive value. Select all the correct statements regarding this function.

(a) Increasing the value of b shifts the sigmoid function to the right (i.e., towards positive infinity)

(b) Increasing the value of b shifts the sigmoid function to the left (i.e., towards negative infinity)

(c) Increasing the value of w decreases the slope of the sigmoid function

(d) Increasing the value of w increases the slope of the sigmoid function

Correct Answer: (b),(d)
Solution: Plot the sigmoid function using graphing tools, keeping w and b as variables. Observe how the slope and y-intercept of the sigmoid function change.

7. You are training a model using the gradient descent algorithm and notice that the loss decreases and then increases after each successive epoch (pass through the data). Which of the following techniques would you employ to enhance the likelihood of the gradient descent algorithm converging? (Here, η refers to the step size.)

(a) Set $\eta = 1$

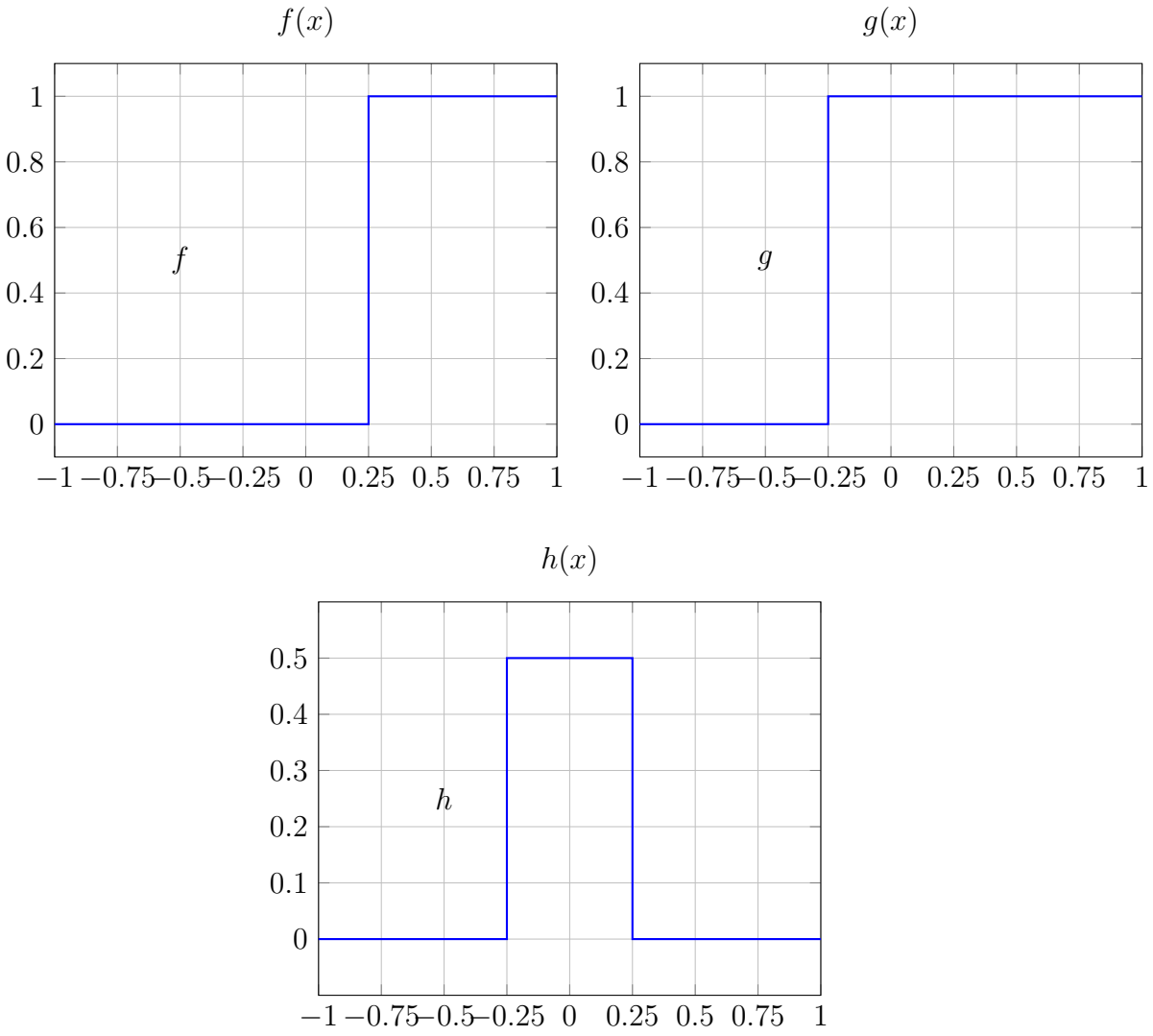
(b) Set $\eta = 0$

(c) Decrease the value of η

(d) Increase the value of η

Correct Answer: (c)
Solution: The loss is oscillating around the minimum, indicating that our η (step size) is too high. Hence, lowering η will increase the likelihood of converging to the minimum.

8. The diagram below shows three functions f, g and h. The function h is obtained by combining the functions f and g. Choose the right combination that generated h.



(a) $h = f - g$

(b) $h = 0.5 * (f + g)$

(c) $h = 0.5 * (f - g)$

(d) $h = 0.5 * (g - f)$

Correct Answer: (d)

Solution:

To verify the solution $h = 0.5 \cdot (g - f)$, we analyze the given graphs:

Observing f: The function f is a sigmoid function centered at $x = 0.25$, transitioning smoothly from 0 to 1 as x increases.

Observing g: The function g is another sigmoid function but shifted to the left, centered approximately at $x = -0.25$. It also transitions smoothly from 0 to 1.

Observing h: The function h exhibits the following characteristics:

- A plateau around $x = 0$ with a constant value of 0.5.
- A transition to 0 outside the overlapping regions of f and g.

Derivation of $h = 0.5 \cdot (g - f)$:

(a) Both f and g are sigmoid functions with different shifts.

(b) The difference $g - f$ is positive where $g > f$, creating the observed plateau effect.

(c) By scaling $g - f$ by 0.5, the difference is normalized to a maximum of 0.5, as seen in the graph for $h(x)$.

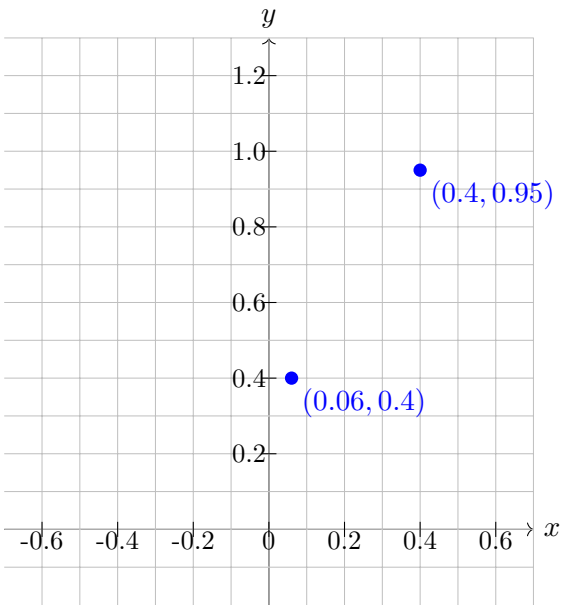
(d) Outside the overlapping regions, $f \approx 0$ or $g \approx 0$, so $h(x)$ trends toward 0, as expected.

Thus, the equation $h = 0.5 \cdot (g - f)$ perfectly describes the observed behavior of $h(x)$.

The function $h(x)$ is correctly given by:

$$h(x) = 0.5 \cdot (g(x) - f(x)).$$

9. Consider the data points as shown in the figure below,



Suppose that the sigmoid function given below is used to fit these data points.

$\frac{1}{1+e^{-(20x+1)}}$

Compute the Mean Square Error (MSE) loss $L(w, b)$

(a) 0

(b) 0.126

(c) 1.23

(d) 1

Correct Answer: (b)

Solution: The given sigmoid function is:

$$f(x) = \frac{1}{1 + e^{-(20x+1)}}$$

and the Mean Square Error (MSE) loss is defined as:

$$L(w, b) = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

Step 1: Data Points The given data points are:

$$(x_1, y_1) = (0.06, 0.4), \quad (x_2, y_2) = (0.4, 0.95)$$

Step 2: Predicted Values Using the sigmoid function:

$$f(x_1) = \frac{1}{1 + e^{-(20 \cdot 0.06 + 1)}} = \frac{1}{1 + e^{-2.2}} \approx 0.9002$$

$$f(x_2) = \frac{1}{1 + e^{-(20 \cdot 0.4 + 1)}} = \frac{1}{1 + e^{-9}} \approx 0.9999$$

Step 3: Squared Errors

$$\text{Error}_1 = (f(x_1) - y_1)^2 = (0.9002 - 0.4)^2 = (0.5002)^2 \approx 0.2502$$

$$\text{Error}_2 = (f(x_2) - y_2)^2 = (0.9999 - 0.95)^2 = (0.0499)^2 \approx 0.0025$$

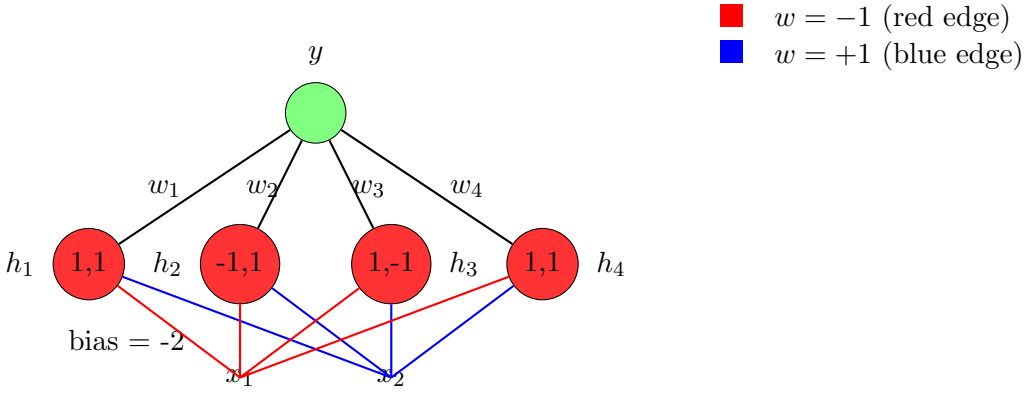
Step 4: Mean Square Error (MSE)

$$L(w, b) = \frac{1}{n} \sum_{i=1}^{n} \text{Error}_i = \frac{1}{2} (0.2502 + 0.0025)$$

$$L(w, b) = \frac{1}{2} \cdot 0.2527 = 0.12635$$

$$L(w, b) \approx 0.12635$$

10. Suppose that we implement the XOR Boolean function using the network shown below. Consider the statement that "A hidden layer with two neurons is suffice to implement XOR". The statement is



(a) True

(b) False

Correct Answer: (a)
Solution:

(a) First, recall the XOR truth table:

x_1	x_2	XOR Output
0	0	0
0	1	1
1	0	1
1	1	0

(b) Looking at the given network structure:

- It has 4 hidden neurons, each receiving inputs with weights -1 (red) or +1 (blue)
- Each hidden neuron shows two values (e.g., 1,1 or -1,1)
- Network has a bias of -2

(c) Key insight for sufficiency of two neurons:

- XOR function requires two line separators in the input space
- Each neuron can act as one line separator
- Two neurons together can create the necessary separation pattern

(d) Mathematical justification:

- First neuron: Can create a line separating (0,0) from (1,1)
- Second neuron: Can create a line separating (0,1) from (1,0)
- The output layer combines these separations to implement XOR

(e) Implementation with minimum neurons:

- Neuron 1: Detects when both inputs are 1
- Neuron 2: Detects when both inputs are 0
- Output layer: Combines these signals to produce correct XOR output

Conclusion: While the given network uses 4 neurons, it's overparameterized for the XOR problem. Two neurons are mathematically sufficient because:

- XOR is not linearly separable (impossible with single neuron)
- Two neurons provide the minimum geometric complexity needed
- Additional neurons (as shown in the network) may aid training but aren't necessary

Therefore, the statement is **True**.